



Home Page



← → 🔍 <https://www.kaggle.com/competitions/sentiment-analysis-on-movie-reviews/leaderboard>

Sentiment Analysis on Movie Reviews

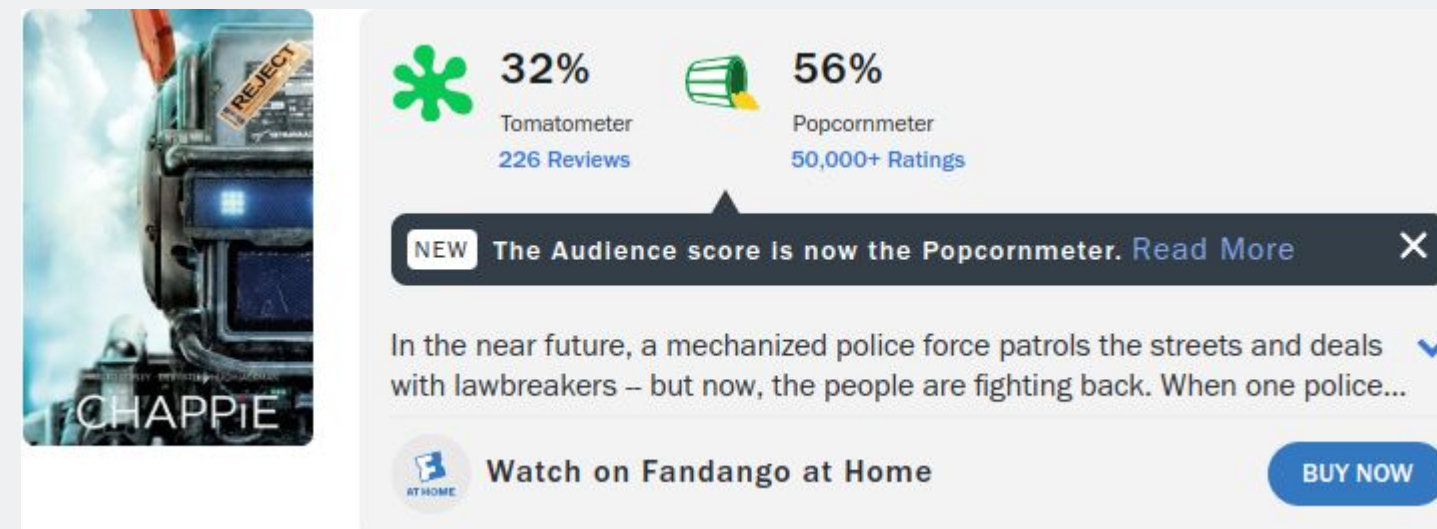
Predictive analytics with machine learning

by Tiago Pedro



Competition Description

For this competition we are delivered a dataset from Rotten Tomatoes, with comments and the sentiment associated with the comment, we have a training data set and a test data set for the submission.



Joshua C

I could not tell you a single thing that happened in the movie. Why was hugh jackman in the movie? Why did daddy even survive, he played no meaningful role in the plot but be dripped o...

Rated 1.5/5 Stars • 07/13/24

[Full Review](#)

Pedro Henrique C

I liked Chappie at first but then everything got bad for good

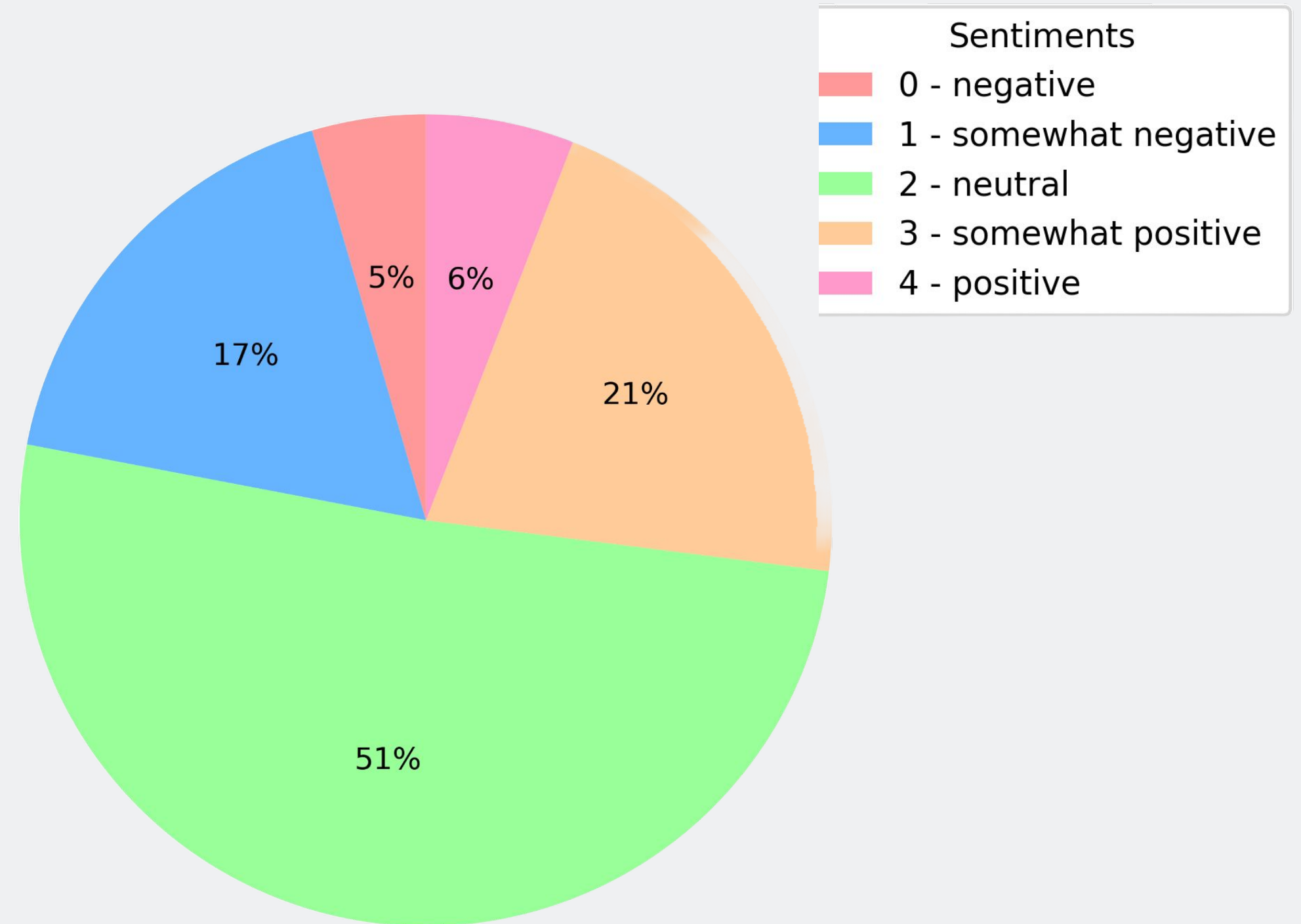
Rated 3/5 Stars • 08/02/24

[Full Review](#)



EDA

The train dataset is composed of 4 columns, 'Phraseld', 'Sentenceld', 'Phrase' and 'Sentiment', it has 156060 rows, with no duplicated or null values, and the sentiment is a integer that ranges from 0 to 4.



[Home Page](#)[Introduction](#)[Project Pipeline](#)

Playground Prediction Competition - Sentiment Analysis on Movie Reviews

EDA

‘Phraseld’

An integer with the id of the phrase, a phrase then is split in multiple sentences

‘SentenceId’

An integer with the number of the sentence

‘Phrase’

The string that we will use to train the model, containing the comment or parts of it

[Home Page](#)[Introduction](#)[Project Pipeline](#)

Playground Prediction Competition - Sentiment Analysis on Movie Reviews

Data Preprocessing



Null values

The data set had no Null values



SMOTE

Because of the high count of Neutral sentiments, we balanced the dataframe with SMOTE



Spacy

To preprocess the text, including tokenization, lemmatization, and removal of stopwords.



TF-IDF

Term Frequency-Inverse Document Frequency, to convert the text into numbers



Model Evaluation & Metrics

Model	Accuracy	MAE	RMSE	R2 score
Logistic Regression	0.57	0.49	0.80	0.66
Random Forest	N/A	0.61	0.88	0.06
SVM	0.65	0.40	0.72	0.66
KNN	0.52	0.59	0.90	0.71
XGBClassifier	0.55	0.53	0.53	0.63



Home Page

Introduction

Project Pipeline



Playground Prediction Competition - Sentiment Analysis on Movie Reviews

Hyperparameter Tuning

Logistic Regression

C (Inverse of Regularization Strength)
penalty (Regularization Type)
solver (Algorithm to Optimize the Model)
max_iter (Maximum Iterations)

SVM

C (Regularization Parameter)
kernel (Kernel Function)
gamma (Kernel Coefficient for 'rbf', 'poly', and 'sigmoid')

[Home Page](#)[Introduction](#)[Project Pipeline](#)

Playground Prediction Competition - Sentiment Analysis on Movie Reviews

Final Model

For time purposes we used the Logistic Regression, with 'C': [10], 'penalty': ['l2'], 'solver': ['liblinear'], as the SVM Hyperparameter Tuning is a long process when started it would not have time or computing power to work with it.

```
param_grid_lr = {  
    'C': [10],  
    'penalty': ['l2'],  
    'solver': ['liblinear']  
}
```




Home Page

Introduction

Project Pipeline



Playground Prediction Competition - Sentiment Analysis on Movie Reviews

Kaggle Submission

YOUR RECENT SUBMISSION



submission.csv

Score: 0.56757



Submitted by Tiago "ShaQ" Pedro · Submitted a minute ago



Jump to your leaderboard position

569

Vagabond



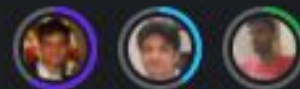
0.56902

4

10y

570

Abhinav Unnam



0.56893

10

10y

571

Md. Mosharaf Hossain



0.56750

2

10y

572

Daniel B.



0.56658

3

10y

[Home Page](#)[Introduction](#)[Project Pipeline](#)[Conclusion](#)

Playground Prediction Competition - Sentiment Analysis on Movie Reviews

Key Insights & Challenges



Machine Learning Models

Choosing the right ones for the project and the ones possible to use



Tokenization

The way we can compile a text to different type of values to input on the ML model



Computing power struggle

Some ML models take longer to train or are even impossible to use with the current computing power



Hyperparameter Tuning

Time consuming to try various ML, the need to test just the right from the 'standart' tests

[Home Page](#)[Introduction](#)[Project Pipeline](#)[Conclusion](#)

Playground Prediction Competition - Sentiment Analysis on Movie Reviews

Possible Improvements

- Use of Transformers-Based models like BERT/DistilBERT
- Use of Deep Learning Models like RNNs or CNNs
- More computing power to be able to use the models with the full data set
- Test with different Tokenizers like KERAS
- Test different word embeddings like GloVe or Word2Vec

[Home Page](#)[Introduction](#)[Project Pipeline](#)[Conclusion](#)

Playground Prediction Competition - Sentiment Analysis on Movie Reviews

Conclusion

Machine learning is not only about the model we chose, but the way we manage the data before training the models, in this case using text to predict a data, there are various ways to treat them with different libraries for Tokenization, the lower score on the competition show the need to study and understand these libraries to optimize the data for better results.



Title Page

Reporters

Introduction

Conclusion

Questions



**Thank you for your time and
attention,ask any questions
you may have**

Project made by Tiago Pedro